



# PySpark Installation Guide

讲师：轩宇



# PySpark

**Python Introduction**

**Installing Python Windows**

**PyCharm Settings**

**PySpark Configuration**

**Quick Start Example**



Welcome to Python.org x

Python Software Foundation [US] | <https://www.python.org>

Python PSF Docs PyPI Jobs Community

python™

Search GO Socialize Sign In

About Downloads Documentation Community Success Stories News Events

```
# Python 3: Simple output (with Unicode)
>>> print("Hello, I'm Python!")
Hello, I'm Python!

# Input, assignment
>>> name = input('What is your name?\n')
>>> print('Hi, %s.' % name)
What is your name?
Python
Hi, Python.
```

**Quick & Easy to Learn**











Experienced programmers in any other language can pick up Python very quickly, and beginners find the clean syntax and indentation structure easy to learn. [Whet your appetite](#) with our Python 3 overview.

1 2 3 4 5

Python is a programming language that lets you work quickly and integrate systems more effectively. [>>> Learn More](#)

Get Started Download Docs Jobs

# THE 2016 TOP PROGRAMMING LANGUAGES

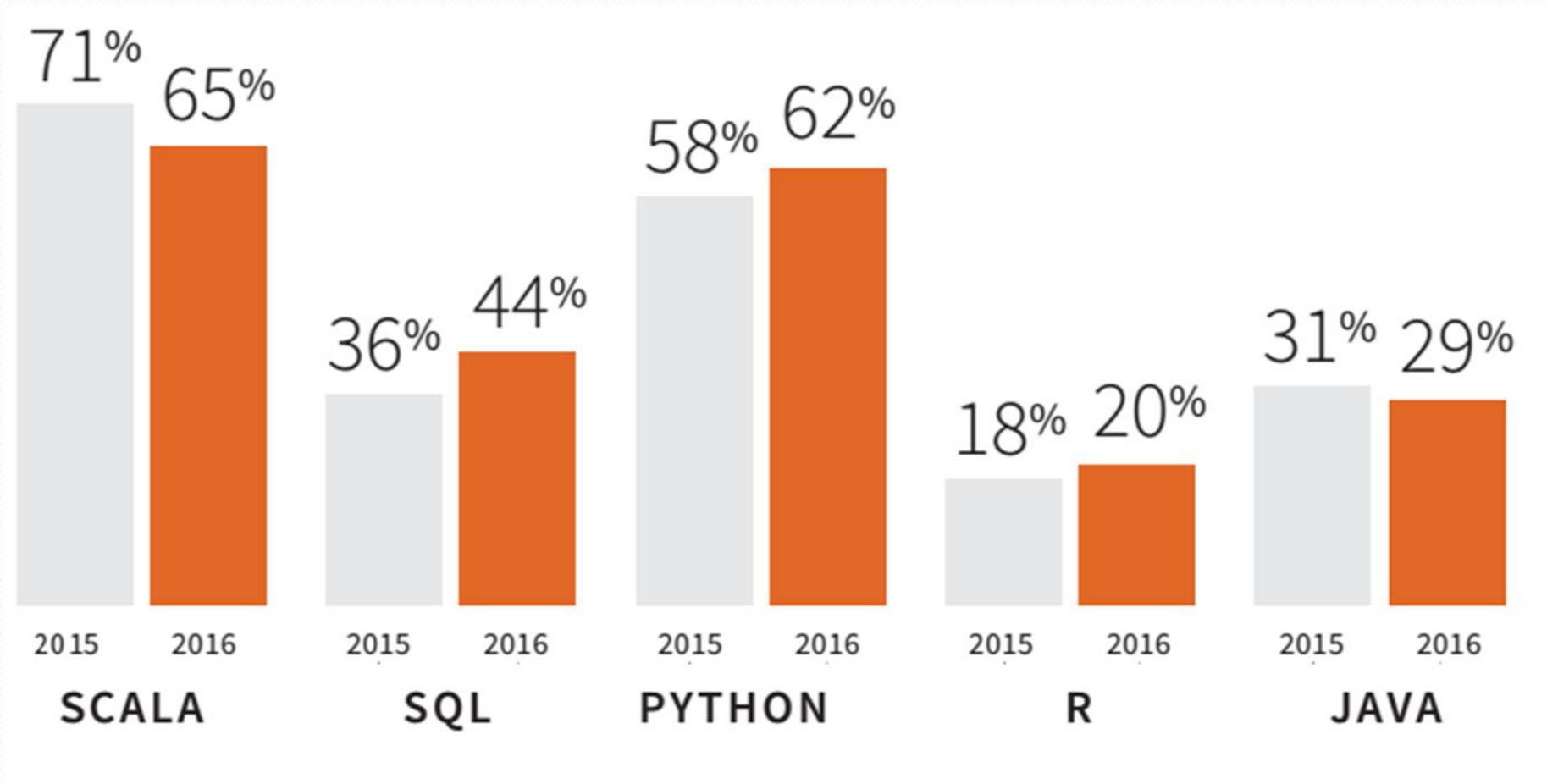
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

# LANGUAGES USED IN APACHE SPARK

---

Respondents were allowed to select more than one language.

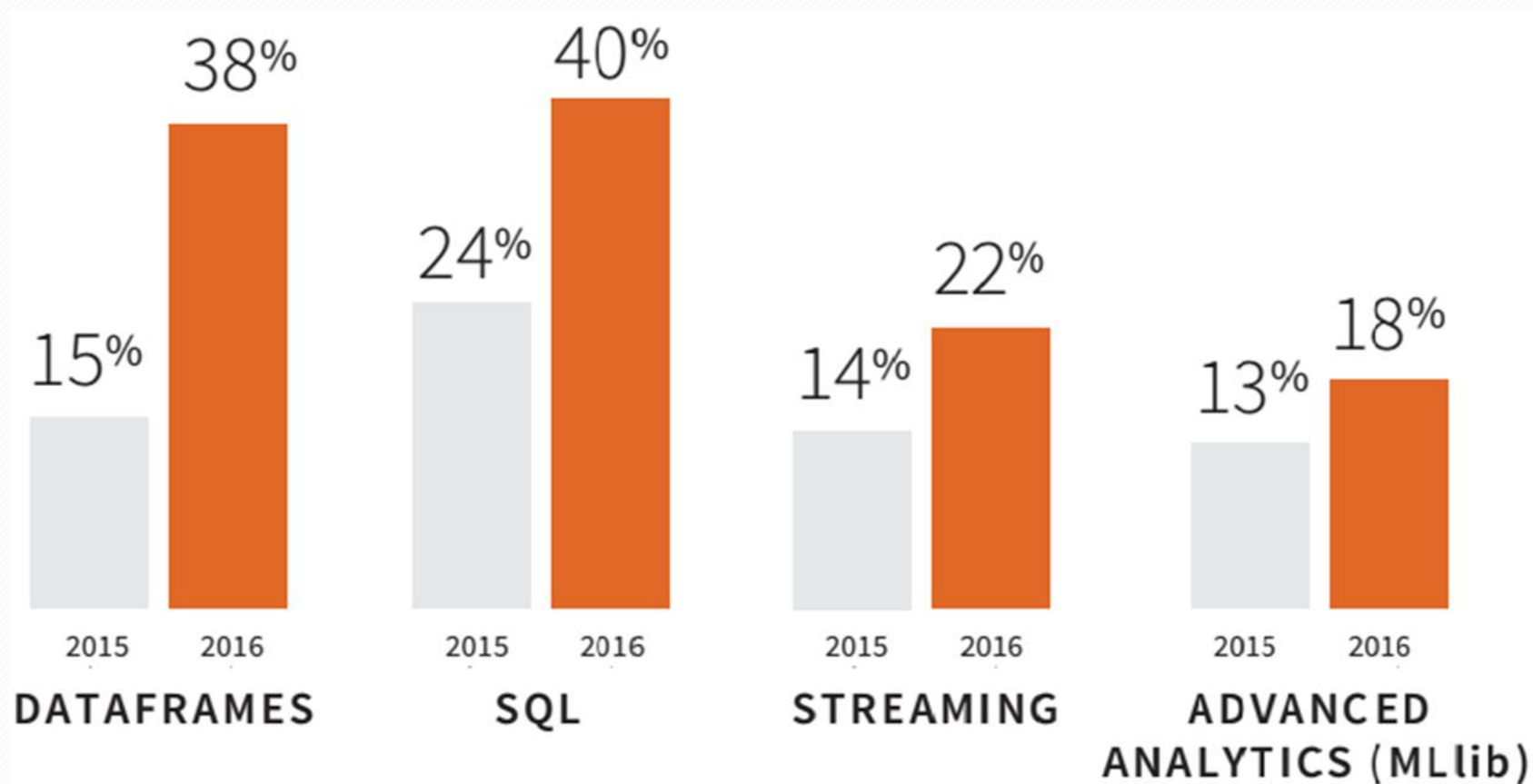




# SPARK COMPONENTS USED IN PRODUCTION

---

Respondents were allowed to select more than one component.



# APACHE SPARK'S FASTEST GROWING AREAS IN 2016

---



DATAFRAME USERS

+153%

SPARK SQL USERS

+67%

SQL



STREAMING USERS

+57%

ADVANCED ANALYTICS  
USERS (MLLIB)

+38%



## Hadoop大数据开发工程师 / 15k-30k

任职要求：

- 1、两年以上hadoop的应用开发经验，至少一个企业级数据仓库项目开发经验
- 2、优秀的编程开发能力，精通Java；
- 3、对数据结构、算法有深刻理解，有预测模型，行为分析模型，推荐模型
- 4、熟悉python shell、perl中的一种；
- 5、熟悉hadoop生态圈中的hive、impala、kafka、flume等，对hive、in
- 6、技术敏感，有一定独立分析，技术研究能力，乐于接受挑战，具有良好

## 大数据Hadoop开发工程师 / 15k-25k

职位要求：

1. 计算机或相关专业本科或以上学历
2. JAVA基础扎实（最好3年以上工作经验）

熟悉Hadoop、Hive，理解云计算，对Hadoop、Hive源码有研究优先，熟悉MapReduce编程，有过大数据处理经验者优先；

hadoop, nutch, hive, hbase, pig, AWS或阿里云, nosql等

3. 熟悉linux环境，熟练使用至少一种脚本语言，bash/perl/python/php/ruby
4. 有互联网公司经验或者对技术有热情的加分
5. 有用户画像项目相关经验加分

工作地址

北京 - 海淀区 - 白石桥 - 西外大街168号腾达大厦29层

[查看地图](#)

## 高级大数据（hadoop）开发工... / 20k-30k

(1)本科及其以上学历，计算机、数学等相关专业；

(2)五年以上工作经验，三到五年大数据项目经验，熟悉Java并能快速找到原因，并有相应的源码读写能力，熟悉分布式系统开发；

(3)深入理解linux操作系统，最好能够独立解决Linux系统问题，能写复杂的shell/python脚本；

(4)熟悉CDH\HDP平台，能独立安装，有kerberos和HA的安装及使用经验，有故障排直能力；

(5)熟练掌握HDFS YARN Hive HBase Spark Storm中一个或多个大数据技术，对其有过深入研究和优化。

## Hadoop大数据开发/架构 / 28k-45k

(1) 建立大规模数据的数据仓库基础架构，并根据上层业务需求和计算逻辑持续优化

包括但不限于：调度系统、元数据管理、数据质量监控、高效数据同步、流式数据

中遇到的计算平台优化、数据处理技术、基础工具使用等技术问题研究大数据产品

专业

ark\Flume等开源技术，有2年以上的实际工作经验，对相关系统源代码有研究

资源管理、调度算法、并行数据处理有自己的理解；

至少一种编程语言，C++、Java、Python、Scala；

经验者优先；

先，例如元数据管理、OLAP引擎、数据同步等；

，有进取心

[查看地图](#)